(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
2 June 2005 (02.06.2005)

PCT

(10) International Publication Number
WO 2005/050474 A2

(51) International Patent Classification[7]: G06F 17/21

(21) International Application Number:
PCT/IB2004/052405

(22) International Filing Date:
12 November 2004 (12.11.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
03104316.9 21 November 2003 (21.11.2003) EP

(71) Applicant *(for DE only)*: **PHILIPS INTELLECTUAL PROPERTY & STANDARDS GMBH** [DE/DE]; Steindamm 94, 20099 Hamburg (DE).

(71) Applicant *(for all designated States except DE, US)*: **KONINKLIJKE PHILIPS ELECTRONICS N. V.** [NL/NL]; Groenewoudseweg 1, NL-5621 BA Eindhoven (NL).

(72) Inventors; and

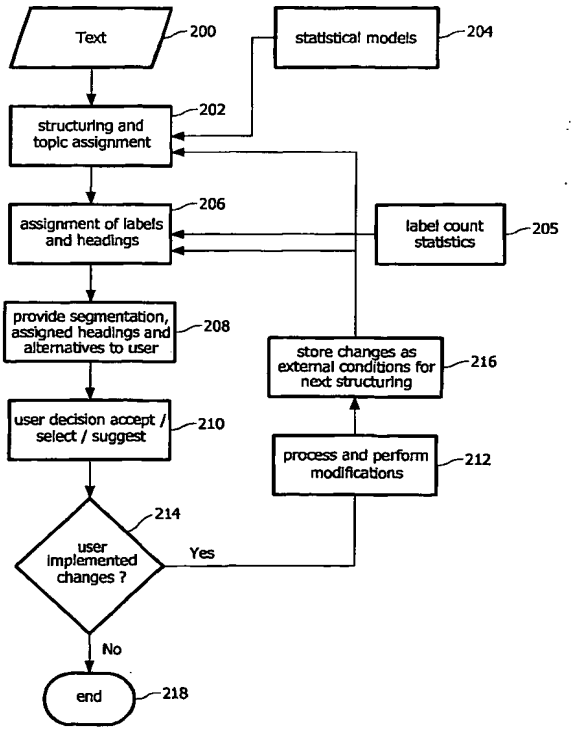
(75) Inventors/Applicants *(for US only)*: **PETERS, Jochen** [DE/DE]; c/o Philips Intellectual Property & Standards GmbH, Weisshausstr. 2, 52066 Aachen (DE). **MATUSOV, Evgeny** [RU/DE]; c/o Philips Intellectual Property & Standards GmbH, Weisshausstr. 2, 52066 Aachen (DE). **MEYER, Carsten** [DE/DE]; c/o Philips Intellectual Property & Standards GmbH, Weisshausstr. 2, 52066 Aachen (DE). **KLAKOW, Dietrich** [DE/DE]; c/o Philips Intellectual Property & Standards GmbH, Weisshausstr. 2, 52066 Aachen (DE).

(74) Agents: **MEYER, Michael** et al.; Philips Intellectual Property & Standards GmbH, Weisshausstr. 2, 52066 Aachen (DE).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD,

[Continued on next page]

(54) Title: TEXT SEGMENTATION AND LABEL ASSIGNMENT WITH USER INTERACTION BY MEANS OF TOPIC SPECIFIC LANGUAGE MODELS AND TOPIC-SPECIFIC LABEL STATISTICS



(57) Abstract: The invention relates to a method, a computer program product, a segmentation system and a user interface for structuring an unstructured text by making use of statistical models trained on annotated training data. The method performs text segmentation into text sections and assigns labels to text sections as section headings. The performed segmentation and assignment is provided to a user for general review. Additionally, alternative segmentations and label assignments are provided to the user being capable to select alternative segmentations and alternative labels as well as to enter a user defined segmentation and user defined label. In response to the modifications introduced by the user, a plurality of different actions are initiated incorporating the re-segmentation and re-labelling of successive parts of the document or the entire document. Furthermore the method comprises a learning functionality, logging and analyzing user introduced modifications for adaptation of the method to the user's preferences and for further training of the statistical models.

WO 2005/050474 A2

MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.